THE RELIABILITY OF HOLISTIC AND ANALYTIC EVALUATIONS OF EFL ESSAYS BY TURKISH UNIVERSITY PREPARATORY STUDENTS

A THESIS
SUBMITTED TO THE FACULTY OF HUMANITIES AND LETTERS
AND THE INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES
OF BİLKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF ARTS
IN THE TEACHING OF ENGLISH AS A FOREIGN LANGUAGE

BY
[redacted]
AUGUST 1998

PE
1068
.T8
S34
1998

THE RELIABILITY OF HOLISTIC AND ANALYTIC EVALUATIONS OF EFL ESSAYS BY TURKISH UNIVERSITY PREPARATORY STUDENTS

A THESIS
SUBMITTED TO THE FACULTY OF LETTERS AND HUMANITIES
AND THE INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF ARTS
IN THE TEACHING OF ENGLISH AS A FOREIGN LANGUAGE

Şehnaz Şahinkarakaş
tarafından Değiştirilmiştir.

BY

ŞEHNAZ ŞAHINKARAKAŞ

AUGUST 1993

PE
1068
.T8
S34
1993
B 13800

to my husband

with my love,

ABSTRACT

Title: The reliability of holistic and analytic evaluations of the EFL essays by Turkish University preparatory students
Author: Şehnaz Şahinkarakaş
Thesis Chairperson: Dr. Dan J. Tannacito, Bilkent University, MA TEFL Program
Thesis Committee Members: Ms. Patricia Brenner, Dr. Linda Laube, Bilkent University, MA TEFL Program

This study attempted to investigate a reliable method of scoring essays. Two hypotheses were tested. Observations were made pertaining to the scoring system used at the preparatory school of Çukurova University. A total of 150 EFL preparatory students participated in the study. These students wrote two essays: one for the first hypothesis and one for the second. The first essays were rated analytically by the teachers at Çukurova University. The second essays were rated holistically and analytically by four raters who have experience at EFL teaching situation for at least five years. Correlations were made to find the relationships between the scores given by the raters for the scoring methods.

The first hypothesis was that the scoring system used at Çukurova University did not have a high level of reliability. The correlational analysis of data rejected this hypothesis (r=.97). However, descriptive analysis showed that the correlation of the scores alone would not be sufficient to claim that this system was reliable. In fact, observations indicate the raters who scored essays for the second time saw the first scores, thus creating a self-fulfilling bias.

The second hypothesis was that holistically scored essays have significantly greater reliability than analytically scored ones in this educational context. The analysis of data was twofold: interrater reliability and intrarater reliability. The correlation for interrater reliability indicated that both scoring systems had high reliabilities. The interrater reliability of holistic scoring method was .85, and of analytic scoring method was .84. The difference is negligible.

Since the analytic scoring method has five categories, the study

investigated the reliability of each category individually as well as the total. The analysis of categories revealed that the reliability of the categories was not as high as the total scores for analytic rating. The interrater reliability was .75 for content, .69 for organization, .80 for vocabulary, .82 for language use, and .71 for mechanics.

The correlations for intrarater reliability showed that there was not a significant difference between the two scoring methods ($p<.01$ for both scoring). The intrarater reliability of holistic scoring ranged from .70 to .85 and of analytic scoring from .65 to .86.

However, the categories scored on the analytic rubric had low intrarater reliabilities. The intrarater reliability ranged from .34 to .83 for content, from .23 to .81 for organization, from .46 to .80 for vocabulary, from .63 to .77 for language use, and from .55 to .80 for mechanics.

We may conclude that holistic scoring is more reliable than analytic scoring. Although the total scores of analytic scoring might have high reliability, the categories of this scoring method might have very low reliability which may raise a question about the reliability of analytic scoring.

BILKENT UNIVERSITY

INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES

MA THESIS EXAMINATION RESULT FORM

August 31, 1993

The examining committee appointed by the
Institute of Economics and Social Sciences for the
thesis examination of the MA TEFL student

Şehnaz Şahinkarakaş

---

has read the thesis of the student.
The committee has decided that the thesis
of the student is satisfactory.

Thesis Title	:	Reliability of holistic and analytic evaluation of the EFL essays of Turkish university preparatory students
Thesis Advisor	:	Dr. Dan J. Tannacito Bilkent University, MA TEFL Program
Committee Members	:	Ms. Patricia Brenner Bilkent University, MA TEFL Program
		Dr. Linda Laube Bilkent University, MA TEFL Program

We certify that we have read this thesis and that in our combined opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts.

Dan J. Tannacito
(Advisor)

Patricia Brenner
(Committee Member)

Linda Laube
(Committee Member)

Approved for the
Institute of Economics and Social Sciences

Ali Karaosmanoğlu
Director
Institute of Economics and Social Sciences

ACKNOWLEDGEMENTS

I am most grateful to my advisor, Dr. Tannacito, for his endless patient and help. He taught me not only how to do research but also how to enjoy doing research. I will always remember him and his recommendations in my future studies.

I would like to thank my thesis committee members, Linda and Patricia, and my tutor, Ruth for their support.

I am very grateful to my classmates Alev, Şadiye, Ali and Adnan, who helped me a lot for my study and all my other classmates for encouraging me whenever I needed.

I would like to thank the administrators, the testing committee, my colleagues and the students at Çukurova University for their help and understanding. Deep thanks to Hülya, Nükhet and Meral who never left me alone.

More than thanks are due to Türkay for her great support and help in my statistical analysis. How could I do statistics without her?

And my husband, my mother, my mother-in-law, my children and Fatma. My thanks are also to you for living with me all through my study.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1 INTRODUCTION

Background of the Problem

EFL teachers increasingly are faced with a need to evaluate writing both for classroom and institutional purposes. Evaluating writing is an important measure of a learner's communication ability. The results can be used for testing students' level of proficiency, for placement purposes, and so forth. The need to evaluate writing has led to the development of several methods of assessment in TEFL.

Two of these methods are: analytic and holistic methods of assessment. Analytic assessment uses a detailed rubric in which a list of features and characteristics of writing to be evaluated are mentioned in full detail. The advantage of this assessment is that it allows the teachers to observe the development of the students in different writing characteristics. For example, analytic method can be used to find if a student is better at organization than language use, or better at vocabulary, or at coherence, and so forth. Carey (1988) describes the advantages of analytic assessment:

> This procedure helps a teacher focus on relevant aspects of students' responses and provides a systematic way to assign a partial credit. Just as important, it allows students to see where they lost points. Using this method, teachers can summarize the group's performance on main components, analyze errors, and use the error analysis to evaluate and revise instruction. (p. 191)

Carey mentions that it is easy for the teachers to understand what problems students face in writing by using analytic method because teachers evaluate writing within different categories, not as a whole unit, and so teachers will have the chance to reexamine the problem parts of writing.

Holistic assessment uses a less detailed rubric, using only a general impression scale. This assessment does not allow teachers to evaluate

writing ability within different categories. Rather, it helps teachers evaluate writing quality as a whole unit. The impression of the teacher when s/he reads an essay once is very important for the evaluation. Mann (1988) states the difference between the analytic assessment and holistic assessment:

> Holistic scoring differs from analytic scoring in a dramatic way. Instead of assessing selected composition features, it responds to student writing as a unit. Developed under the auspices of Educational Testing Service, holistic scoring is quick and practical as well as cost effective. Most important, however, is the fact that this scoring option permits the rating of total effectiveness of the writing sample, not just of certain features. (p. 6)

Mann views holistic scoring as a better method of assessment than analytic scoring because the rater will have a chance to evaluate writing as a whole and because holistic is more practical.

These two methods of assessment have advantages and disadvantages. Some researchers think that holistic is a more reliable method to evaluate quality of writing. On the other hand, some researchers think that analytic is a more reliable method. Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) explains reliability as the extent to which a test yields consistent results, i.e., are the test scores precise, stable, and dependable? The reliability can be tested in several ways; two of them -- interrater and intrarater reliability -- will be taken into consideration in this study. Interrater reliability is a way to test how consistent two or more raters are in rating the same writing sample. Intrarater reliability is another way to test how consistent one rater is in scoring the same writing sample twice with a specific time interval between the two rating.

To determine the reliability of any assessment method to be used is of great importance. The reliability of a method, whether analytic or holis-

tic, can be improved through a careful training of raters. If raters are untrained they score essays inconsistently.

Purpose of the Study

The first goal of this research was to investigate the reliability of the method of assessment used at Çukurova University Preparatory School. In this institution approximately ten thousand essays are rated a year. The questions thus arise: What is the consistency, that is, the agreement of the grades given to each essay by different raters, of the current analytic assessment system used in this institution? If there is an inconsistency, what is/are the source(s) of this inconsistency? Is it the rubric itself? Is it the lack of training? Is it the background of the raters? Is it the conditions for evaluation? Is it the process of assessment? How long does the rating procedure last?

Moreover, in order to determine the best method of assessment, whether analytic or holistic, a second question was investigated: Do holistically scored essays or analytically scored essays have significantly greater reliability? To answer this question this study used two different rubrics: the Test of Written English (TWE) (Boyd, 1990) as representative of holistic scoring (see Appendix C) and the ESL Composition Profile (Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey, 1981) as representative of analytic scoring (see Appendix D).

This study tested two hypotheses. The first hypothesis was that the analytic scoring method used at Çukurova University Preparatory School did not have high reliability. The second hypothesis was that holistically scored essays had significantly greater reliability than analytically scored essays.

Significance of the Study

The present study is significant for the institutions that are administering EFL writing exams. This study explicates the reliability of two methods of writing assessment: holistic and analytic methods. It also explains which of these two methods is less time-consuming and suitable to apply.

The study is also significant for TEFL researchers because among the studies that have been done comparing reliability of holistic and analytic scoring, different results have been found. Some research claims that holistic scoring has greater reliability than analytic while other research claims analytic scoring is more reliable than holistic. Therefore, these two scoring methods will be reevaluated to see how they apply to Turkish University preparatory departments.

In sum, this research is significant for institutions and administrators who want to find out the most reliable way of assessing their students' writing, and for the field researchers who want to examine different ways of assessment.

Delimitations and Limitations of the Study

The delimitation of this study was that the data was collected for only undergraduate advanced level students. In the preparatory department of the university, where the researcher collected data, there are about one thousand students at four levels: beginner, lower-intermediate, upper-intermediate and advanced. These levels are divided into two: graduate and undergraduate students. The data consisted of the essays written by only undergraduate advanced level students. It is possible to generalize the results of those students to all levels in this situation since all level students are taking the same type of writing exam, that is, direct writing. The difference is that the beginner and intermediate level students are asked to write short paragraphs or guided compositions.

Despite the difference, the writing section in exams is evaluated with the same type of scoring method (a kind of analytic assessment). Therefore, it is possible to apply the most reliable scoring system which was investigated in this study to all levels in this institution.

The main limitation of this study was the number of writing samples rated. Although the researcher collected 150 essays, this number had to be reduced to 50 because of the time limitation. Since this research was to be done in a short time, it would be very hard to expect the raters to score all 150 essays without fatigue. The data were limited to fifty essays in order to avoid fatigue.

The other limitation of this study was that the participants could not write a second equivalent essay to compare holistic and analytic scoring methods. It would be too burdensome on the administration to let the researcher collect more data since students wrote the essays during their class time. Furthermore, it would be a burden on the raters as well to rate fifty more essays in a limited time.

Conceptual Definitions

There are several definitions for holistic and analytic scoring systems in the field of composition evaluation. Among these, the researcher used a definition by Katz (1988) since she best describes how the guidelines of TWE scoring method (used in this study) can be used to respond student writing as a unit. She explains "even though each level of the guidelines describes specific, required abilities, evaluation requires raters to take all of these parts and blend them into a whole score" (p. 199).

The definition for analytic scoring is that of Hughey, Wormuth, Hartfiel, and Jacobs (1983) since their definition best explains how the ESL Profile (used in this study) can be used to respond student writing as separate features. They describe analytic scoring as "providing a side view, an outline, of an ESL writer's success at composing or synthesizing

the main elements of writing into a connected, coherent, effective piece of written discourse" (p. 139).

On the interrater and intrarater reliability, the researcher used the definitions by Carlson and Bridgeman (1986). They define interrater reliability as a way to test how consistent two or more raters are in rating the same writing sample and intrarater reliability as a way to test how consistent one rater is in scoring the same writing sample twice with a specific time interval between the two rating.

CHAPTER 2 LITERATURE REVIEW

Testing Writing

Testing is an important issue both in foreign language and in second language education because tests are used for a variety of purposes. The word "test" is highly valued in the society according to Lloyd-Jones (1987). He points out that the social goal of universal education has made the numbers of students so large that management requires proof of accomplishment.

However, testing is difficult. It is especially difficult for writing teachers because there are various modes and styles which cause complexity in judgement. The main reason for this complexity is that everyone learns writing throughout their life, mostly out-of-school, and from different people, and in different style. In addition, Lloyd-Jones mentions another difficulty with the testing of writing: "It represents an effort to record quantitatively the quality of the writing or writing skills of a group of people so that administrators can make policies about educational progress" (p. 155).

Teachers can reduce the problems to a minimum by learning enough about the arts of testing. For example, writing teachers can learn enough about the types of testing writing to apply the appropriate type to their situation. They can also learn enough about the reliability of the scoring method they are using.

Accordingly, it is important for teachers of writing to know about the ways to test writing, the scoring methods and the reliability of these scoring methods.

Types of Writing Tests

Testing writing has two kinds of measurement: Indirect and direct tests of writing. An indirect test asks students to respond to questions about composition (Jacobs, Zinkgraf, Wormuth, Hartfield and Hughey, 1981) often

in a multiple choice format (Carlson & Bridgeman, 1986). Although indirect tests of writing are commonly referred to as objective, Hamp-Lyons (1990) thinks this is a misnomer, since human judgement is still central while creating the array of questions and possible answers. Hamp-Lyons describes an indirect measure of writing by saying that:

> It does not require the test taker to write continuous prose, although she or he may write some words, and there is no room for personal interpretation by the test taker since possible answers are provided and the "correct" one already decided upon. (p. 6)

Indirect measures of writing became less popular in the 1970s after emphasis on language as communication which calls for direct test of writing increased. A direct test of writing has at least the following five characteristics, according to Hamp-Lyons:

1. Each individual actually, physically writes at least one piece of continuous text.
2. While the writer is provided with a set of instructions and material, s/he is given a considerable room within which to create a response to the prompt.
3. Each written text is read by at least one, usually more, human reader-judges who has been through some preparation or training for evaluation process.
4. Each judgement made by readers are tied to some common standard measurement, such as a description of expected performance at certain levels or one or several rating scales.
5. Readers' responses to the writing are expressed as a number or numbers of some kind, not written or verbal comments.

With the new approach to language as communication, according to Carlson & Bridgeman (1986), direct measures of writing started serving as the preferred means for assessing writing performance because they more nearly

approximate real discourse. Further, they report that writing samples permit the evaluation of aspects of writing, such as organization, coherence, and the elaboration of ideas which are not measured with indirect measures. According to Jacobs at al. (1981), the benefits of a direct test of writing are that it:

1. emphasizes to learners the importance of language for communication;
2. promotes a closer match between what is taught and what is tested;
3. is more valid;
4. is easier to prepare;
5. produces more meaningful and interpretable results;
6. can indicate level of proficiency and strengths and weaknesses in the writing skill;
7. can be highly reliable if properly administered and evaluated;
8. utilizes the important intuitive, albeit subjective, resources of other participants in the communication process-the readers of written discourse.

Scoring Writing Methods

The importance of direct tests of writing forced the search for reliable and valid scoring methods. Carlson & Bridgeman mention the need to change the current scoring methods used with indirect tests when they say:

> With the development of competence in basic communication skills (writing, speaking listening and reading) as a primary goal for education and with the recognition that many students pass through our educational system with inadequate English-language competence, educators are reappraising their methods and redefining their objectives. (p. 126)

The search for direct writing assessment brought two main scoring methods: holistic and analytic. The former evaluates writing as a whole whereas the latter evaluates different features of writing, such as coherence, vocabulary, and language use.

Holistic Scoring

Holistic scoring responds to student writing as a whole discourse. The raters are trained on a set of instructions, called a rubric, to guide their rating. The TOEFL Test of Written English (TWE) uses a holistic scoring method with a rubric scored from 1 to 6 (see Appendix C).

Holistic scoring has been the most popular assessment tool in writing. As Huot (1990) mentions "many scholars see it as the major means of direct writing evaluation. Others contend that holistic scoring has proven to be the best economical, flexible and applicable of the direct writing instruments" (p. 201).

Gregory (1991) lists six reasons for the popularity of holistic scoring:

1. Low cost, especially if compared with multiple choice type of scoring. The biggest expense will be to raters but since most projects are brief it will not be very expensive;
2. The efficiency of test administration: Tests can be administered in a 45-50 minute class period;
3. High reliability (over .85);
4. The appeal of a holistic approach is to see things as units, as complete, and as wholes;
5. Holistic reading is thought to be face-to-face encounter because the writer's mind embodied in written expression and reader's mind attempting to see what is being communicated;
6. Score descriptions do not vary from year to year. A paper earning a 6 in 1990 should earn 6 in 1993.

Besides these advantages "holistic scoring method has the advantage of being very rapid" (Hughes, 1989, p. 86). Hughes states that an experienced rater can score a one-page piece of essay in just a couple of minutes or even less. There are some other researchers who agree with him (Carlson &

Bridgeman, 1986; Carlson, Bridgeman, Camp & Waanders, 1985; Cooper, 1977; Gregory, 1991; and Mann, 1988). They all agree that the holistic rating procedure rarely requires more than two minutes per paper. On the other hand, since in the holistic scoring method "the score must represent what a sophisticated reader interprets as a total effect" (Lloyd-Jones, 1987, p.164), it is believed to be impressionistic and hence by some researchers, unreliable.

Huot (1990), for example, reexamining the reliability and especially validity of holistic scoring, points out that holistic scoring is at a distinct disadvantage because it is an individually scored test. According to him, the scores must be generalized to show holistic scoring results reflecting writing quality. Huot states that:

> the more reliable a test, the more we can generalize about its outcomes. . . . In other words, we must be able to generalize scores if we wish to claim that holistic scoring results reflect writing quality and ability. So, the ability to generalize about scores received from holistic rating procedures is limited due to its low reliability. (p. 203)

In a study comparing scoring techniques, Perkins (1983) mentions that published research on holistic scoring in terms of reliability and concurrent validity has yielded contradictory findings. He reports the results of a study made by Diederich (1974, as discussed in Perkins, 1983) which show that "out of the 300 essays graded, 101 received every grade from 1 to 9, 94% received either seven, eight or nine different grades; and no essay received less than five different grades from 53 readers" (p. 653).

Fortunately, it is possible to reach reasonably high reliability with holistic scoring when the following points are considered. Thorough training of the readers is necessary. Carlson & Bridgeman (1986) believe that the problem that comes out with the disparity of the students' expect-

ed skills can be resolved by training and reaching a consensus about how to evaluate such essays. Carlson & Bridgeman report high reliability (.80 to .85) for holistic scoring after training, in a study with native student population.

Additionally, holistic scoring can be highly reliable when raters from similar backgrounds are carefully trained. Cooper (1977) believes that in addition to training, raters background is also important. He states that it is possible to improve reliability from a range of .30 to .75 before training to a range of .73 to .98 after training. The following statement best describes what he believes about the reliability of holistic scoring:

> When raters are from similar backgrounds and when they are trained with a holistic scoring guide--either one they borrow or devise for themselves of the spot--they can achieve nearly perfect agreement in choosing the better of a pair of essays; and they achieve scoring reliabilities in the high eighties and low nineties on their summed scores from multiple pieces of a student's writing. (p. 19)

In a study that examines TOEFL Test of Written English (TWE), Boyd (1990) points out the importance of training for reliability of holistic scoring. She states in her final analysis that "it is not the Scoring Guide that guarantees reliable scoring of TWE papers, but the nature of the training that the readers receive" (p. 101).

Mitchell and Anderson (1986) report the high reliability of holistic scoring they used in a study about the reliability of holistic scoring. Each essay was read by two raters and if the papers received more than one score disagreement, a third reading was needed. Mitchell and Anderson point out that "a third reading was required on 5.3% of the papers" (p. 772). The reliability was .94.

Homburg (1984) also mentions the effect of training on the reliability of holistic scoring in a study that discusses the relationship between

subjective evaluation and objective measures of ESL writing proficiency. He concludes that "holistic evaluation of ESL compositions, with training to familiarize readers with the types of features present in ESL compositions, can be considered to be adequately reliable and valid" (p. 103).

Gregory (1991) in his study examining the weaknesses of holistic assessment accepts the high reliability (.80 or above) of holistic scoring. To mention his belief about the reliability, he quotes Hogan and Mishler (1981) who say that "most researchers agree that this level of reliability (over .80) is possible, despite a widespread notion to the contrary among laypersons" (p. 7).

Nevertheless, we can find some studies that show low reliability for the holistic method with trained raters. Vaughan (1990), for example, conducted a study in which she evaluated the "process" by which raters make their decisions during holistic assessment. In her research the passing essays received 4, 5, or 6 and failing essays received 3, 2, or 1. The raters were trained and asked to grade six essays commenting verbally in a think-aloud procedure as they read. These essays were rated beforehand at the university by teachers of that university. Vaughan reports that "the original raters (teachers of the university) passed only two of the six essays (33%). On the other hand, the raters in this study awarded essays a passing grade 57 percent of the time" (p. 115). Vaughan's study shows low reliability of holistic scoring method. The essays that were rated by the original teachers of the university received 33% success whereas the raters in Vaughan's study passed 57% of the same essays.

Holistic scoring has some other disadvantages besides the question of low reliability of the scoring method. Mann (1988) states that the use of holistic scoring is inappropriate in teaching situations where diagnostic data is required since it gives only a general rating.

Yet, Gregory (1991) sees this as an advantage. He points out that

"language is not learned from subset to subset, structure to structure. Instead, it is under normal conditions 'learned and used all of a piece' so to speak in a holistic fashion" (p. 10).

Therefore, it is possible to find various results for the reliability and the advantages of holistic scoring but the advantages seem to outweigh the disadvantages.

Analytic Scoring

The analytic scoring method is another method of direct writing assessment which provides separate scores. Perkins (1983) defines this procedure as it involves the separation of the various features of a composition into components for scoring purposes. This definition can be applied to the ESL Composition Profile (see Appendix D) as an analytic scoring method. Although the authors consider the ESL Profile as a holistic scoring method (Jacobs et al., p. 29), it fulfills our definition of an analytic method by separating scoring for content, organization, vocabulary, language use, and mechanics.

Like holistic scoring, analytic scoring has some advantages and disadvantages. It is believed to permit reliable scores among raters by some researchers (Carey, 1988; Connor, 1990; Hughes, 1989; Jacobs et al, 1981; Perkins, 1983). Connor, for example, reports the reliability of analytic scoring between .81 to .91. Another advantage of analytic scoring is that as Carey states: "it provides a systematic way to assign partial credit. Just as important, it allows students to see where they lost points" (p. 191).

However, Carlson and Bridgeman see this advantage of analytic scoring as an illusion stating that "the reader's general impression is likely to influence ratings on each of the separate aspects being evaluated" (p. 145).

Edward White (1985, cited in Mann, 1988) states the problem with

analytic scales in the following way:

> There is no evidence that writing quality is the result of the accumulation of a series of subskills. To the contrary, the lack of agreement of subskills in the profession suggests that writing remains more than the sum of its parts and that the analytic theory that seeks to define and add up the subskills is fundamentally flawed. Analytic scoring is uneconomical, unreliable, pedagogically uncertain or destructive and theoretically bankrupt. (p. 6)

White thinks that the analytic scoring method has lost its importance because it evaluates writing as subskills and because it is neither practical for testing proficiency nor reliable.

The main disadvantage of analytic scoring is that it is time-consuming. Stephen Wiseman (cited in Carlson and Bridgeman, 1986) found that "four general impression markings (holistic marking) were equivalent in time and effort to one analytic marking" (p. 145). Therefore, rating large group of essays with analytic scoring could cause fatigue in raters and consequently might affect its reliability.

In studies comparing holistic and analytic scoring methods, both have shown some reliable results. Canale, Frenette and Bélanger (1988), in a study evaluating student writing in first and second language, find reliability for holistic scoring quite high (ranging from .83 to .92) and generally high for analytic (ranging from .59 to 90). However, the range of analytic scores was wider than of holistic sores. Carlson, Bridgeman, Camp and Waanders (1985) also report high reliabilities for both scoring methods: from .80 -.85 for holistic and from .80 -.84 for analytic.

Reliability

Validity is the major concern in large-scale testing programs. Stansfield and Ross (1988) defines validity as "the interferences made about a test score: i.e., the degree to which it is useful as a measure of a

particular trait for a particular purpose and for a particular examinee" (p. 4).

Reliability is another concern in such testing programs. "Reliability refers to the capacity of the assessment procedures to rank-order the same samples of writing performance consistently in the same way". (Henning, 1991) Henning means a reliable test yields the same scores or rank-ordering for the same group of students under different conditions.

Stansfield and Ross also state that in essay testing, reliability is of greater than normal importance because essay tests exhibit a good deal of face validity. The face validity of essay tests is that they require the examinee to perform instead of demonstrating knowledge about how to perform. However, it is not possible to say the same for the reliability of essay tests since they are open to sources of error that are not present in multiple-choice tests. Therefore, in direct measures, such as essay tests, reliability becomes an extensive concern.

The acceptable level of reliability differs according to different uses of the test. Cooper (1977) states that "a reliability coefficient of .80 is considered high enough for program evaluation, while a reliability coefficient of .90 for individual growth measurement in teaching and research" (p. 18). Similarly, Jacobs, et al. reports that "reliability coefficients of .85 are usually considered adequate for tests used for placement purposes, but higher reliabilities -- in the nineties -- are desirable and requested for tests which will be the basis for decisions about individuals" (p. 69). The reliability of .80 was accepted high enough in this study.

Interrater Reliability

Reliability can be measured in a number of ways: interrater reliability, intrarater reliability, test-retest reliability, split-half reliability, intertopic reliability. This research will focus on only two of these

reliability measurements: interrater and intrarater reliability.

Perhaps the most common method for assessing the reliability of writing-sample tests is to determine the interrater reliability. (Greenberg, 1986; Hatch & Lazaraton, 1991; Henning, 1991; Lauer & Asher, 1988) Interrater reliability estimates the extent to which two or more raters agree on the score that should be assigned to an essay. That is, two raters will give the same scores for the same essays.

Interrater reliability involves determining the intercorrelation of two or more raters for the same writing sample, and then adjusting the obtained coefficients by use of the Spearman-Brown Prophecy formula, to reflect the average rating of the raters.

The number of the raters is important in determining the interrater reliability because students' final scores are the combination or average of the ratings. Hence, the more raters, the higher the reliability. Rater training is also important for high interrater reliability as well. If the raters are trained carefully, it is possible to achieve high interrater reliability.

Intrarater Reliability

Intrarater reliability indicates how consistent a single rater is in scoring the same set of essays twice with a specified time interval between the first and second scoring. That is, any particular rater would give the same score on both ratings.

The intercorrelation of the two scoring of one rater for the same writing samples is determined and then the coefficient is adjusted to the Spearman-Brown formula to reflect the intrarater reliability.

As it is in interrater reliability, rater training is important to increase intrarater reliability. If the raters are trained well in each scoring session, it is possible to achieve high intrarater reliability.

Conclusion

This review of literature tells how various researchers have viewed the two main scoring methods of testing writing, holistic and analytic, which this study has examined. The two methods have both advantages and disadvantages. Each of the two methods has been found reliable by some researchers, unreliable by others.

This study examined the reliability of these two scoring methods in EFL context in Turkish University Preparatory schools, focusing only the interrater and intrarater reliability.

CHAPTER 3 METHODOLOGY

Introduction

This study aims to determine which is the more reliable way of scoring EFL writing samples in the context of Turkish universities. The scoring method used in the preparatory department of Çukurova University (ÇU) -- a variant of analytic scoring -- was tested using a simple correlational statistical technique in order to examine the reliability of the method at this university. In addition, two distinct scoring methods, analytic (Composition Profile Scale) and holistic (TWE) scoring methods, were correlated to test whether analytic or holistic scoring is more reliable in this setting.

Research Design

I used a focused descriptive combined with a correlational design in order to test the reliability of the scoring method that is used in Çukurova university and to determine whether analytic or holistic scoring method is more reliable. The study is called a focused descriptive because in such studies the researcher observes the activities and take notes (e. g., observational studies), and narrows the scope of the study to a particular set of variables (Larsen-Freeman and Long, 1991). In this study I narrowed the study to the system for scoring writing samples at the advanced level. I used notes obtained during observation to describe how the scoring system took place at Çukurova University. The number of raters, the number of the papers each rater rates, the scoring method, the raters' training, as well as the procedures for spot-checking were described. This qualitative data was used to tell whether the scoring procedure strengthened or weakened the reliability of the scoring system in this institution.

This study is termed correlational as well because correlational studies try to establish a relationship between scaled or scored data on one variable with those on another (Hatch & Lazaraton, 1991). In this

study, I correlated the scores given by the raters for the two scoring methods. This quantitative data examined the relationship between these scores to find the reliability of the scoring methods.

The following sections deal with who the subjects are, how the data were correlated, and what steps were followed.

Sources of data

The population of this study is about 1000 preparatory EFL students at four proficiency levels: beginner, lower-intermediate, upper-intermediate and advanced level of students. A student's proficiency level is determined by a standardized placement test, called The Proficiency Examination prepared by Cambridge University.

At ÇU preparatory school, students study English intensively, i..e., five days a week. Students study the four skills 25 hours per week. At the beginner, lower-intermediate and upper-intermediate levels, 3 hours out of 25 are devoted to the writing skills and at advanced level 4 hours .

The sample for this study is 150 students who are at the advanced level. Students at this level were selected since they were able to write essays rather than short paragraphs. The age and sex of the students were not considered as moderator variables because these variable were not thought to affect the reliability of the scoring methods. This sample (150 essays) was used to test the reliability of scoring method used in this university.

Fifty out of 150 students were selected by a simple random sampling procedure. I prepared 150 pieces of papers. I marked 50 of these papers and put them together. On the other hand, I numbered the essays. I picked one of the pieces for each numbered essay. If the piece was marked, the essay was selected. This sample (50 essays) was used to test whether analytic or holistic scoring method is more reliable.

This study consists of interval data -- the marks given by raters for

holistic and analytic scoring. Interval data tell us how much of a variable to attribute to a person, text, or object precisely. The intervals of measurement can be described. Each interval unit has the same value so that units can be added or subtracted (Hatch & Lazaraton, 1991).

Measurements

This study has two kinds of variables, independent and dependent. An independent variable is a variable that may relate to or influence the dependent variable. Therefore, the independent variables in this study are the raters since they may affect the other variables. A dependent variable is the one that will be influenced by other variables. Therefore, the dependent variables in this study are the scores given by the raters for each essay using analytic and holistic rubric.

I correlated the scores given by raters and calculated both interrater and intrarater reliability in order to determine whether analytic or holistic scoring method is more reliable. Hence, a correlation was applied to holistic and analytic scoring methods separately. All essays were rated twice for each of the scoring methods. I correlated the scores each rater gave in the first and the second ratings because this study also sought to examine how high the intrarater reliability was for each scoring method.

Instruments

This study has two instruments: Holistic and analytic rubrics. The holistic rubric is the one used in the TOEFL Test of Written English (TWE) (see Appendix C). This rubric is scaled from 1 to 6 and each of the six bands is illustrated with four or five descriptors. The descriptors focus on the degree to which the examinee's writing demonstrates rhetorical and syntactic competence. For example, when raters have to make decisions about the specific characteristics of a competent writer, the rubrics beneath each descriptor are designed to assist readers in their assessment.

The analytic rubric is the ESL Composition Profile (see Appendix D). This rubric contains five component scales (content, organization, vocabulary, language use, and mechanics), each focusing on an aspect of composition and weighted according to its estimated significance for effective written communication. The scores used in Jacobs et al. ranged from 34 to 100 originally. However, these scores were reduced by half, ranging from 17 to 50, in order to achieve more reliable results. The total weight for each component is further broken down into numerical ranges that correspond to four mastery levels (excellent to very good, good to average, fair to poor, and very poor). These levels are characterized and distinguished by key-word descriptors which serve as reminders of specific criteria for excellence and of larger concepts in composition.

Procedure

The first step in this study was to observe the testing and scoring procedure at Çukurova University as a non-participant observer. I was allowed by the administrators to observe the testing and scoring procedure because I would describe why the scoring was reliable or not and because I would recommend to the administrators a reliable method of scoring if the one they were using was found to be unreliable.

Observation of the Scoring Method at Çukurova University

I divided the observation into four parts: during the test, before paper rating started, during rating and during the spot-check. I took notes during my observations.

During the test, I observed the attitude of the teachers in the testing room, the amount of time that was devoted to the writing section in the test, and the materials students and the teachers had.

During the pre-rating part, I observed how the teachers were organized to rate the writing section, how many raters were used in rating, how many papers each rater was given to rate, the criteria distributed to the

raters, if they were the teachers of writing skill only or if they were the teachers of any skill.

During the rating part, I observed how the table leader trained the raters on the rubric, the practice rating, the attitude of the raters toward the rubric and scoring and the average time each rater spent per essay.

During the spot-check part, the testing committee randomly selected 25% of the papers in each set to rerate. I observed how the testing committee conducted the spot-check, what they did when they met a discrepancy in scores, and who did the second rating in the case of discrepancies.

The observations took three days: one day when the exam took place, the following day when the rating took place and a third day when the spot-check occurred.

Data Collection

After the observations, I collected the scores given on each essay by the teachers in this institution in order to test the reliability of the raters. There were 150 advanced students who took the exam.

With the consent of the administration and the students (see Appendix A), one day after the exam took place, I gave the same students a topic and asked them to write an essay on this topic in the same mode and the same period of time as they had done in the exam the previous day.

I collected these essays and by random sampling I selected 50 out of 150 essays to test the more reliable way of scoring by using holistic and analytic scoring methods. Since these essays were to be rated with the two scoring methods, I formed a group of four experienced EFL Turkish teachers as raters. The four teachers had at least five years of teaching experience in the Preparatory departments of Turkish universities. These raters met four times to rate the papers, twice for holistic rating and twice for analytic rating.

Rating Procedure

In the first meeting I trained the raters in the holistic scoring method. The training session took about 1.5 hours. They rated each essay with the holistic scoring method. I divided 50 essays into four sets, each of which consisting of 12 or 13 essays. On each essay there were 4 strips of papers each of which had the number of the essay and the name of the rater. The raters rated the group of papers with their name on the top of the score strips. They wrote the score of the essays on these strips. When a rater finished rating one set of essays, I took the top strip off and gave these papers to the other rater whose name was on the top of the strips. The approximate average time for rating each essay was between 1 and 1.5 minutes. The duration of the rating session took less than 1.5 hours. At the end of this rating session, 200 holistic scores were collected: 50 scores from each of four raters.

A week after the first holistic scoring, the same four raters met again to rate the same 50 essays analytically. The raters were trained in the use of the analytic rubric. This training took more than two hours because the average rating time per essay was almost 3-4 minutes. As a result, the rating was continued in a second session (two days later) in order to avoid fatigue. During the first analytic scoring session, the raters judged a few sample essays to help recall their training. I followed the same procedure as in the holistic rating for the organization of the essays. But in the analytic session, I added the name of the analytic categories of the scoring. They wrote the score for each category next to the total score. This procedure was intended to help determine the relationship between the raters for categories as well as total scores. At the end of this session, there were 200 scores for the first analytic session. This rating session lasted about 3.5 hours.

The rating sessions for the two scoring methods took place twice in

order to determine the intrarater reliability. I gave a one month interval between the two ratings for each method so that the raters could not recall the essays and their scoring. One month after the holistic scoring, the raters met again to rate the same 50 papers using the same holistic rubric. I followed the same procedure in the distribution of the essays. This training and rating session lasted 2 hours. At the end of this session, there were 200 scores for the second holistic scoring.

One month after the analytic scoring, the raters met for the last time to rate the same essays using the same analytic rubric. This training and rating session lasted 3.5 hours. At the end of this session, there were 200 scores for the second analytic scoring.

I formed tables to record the data from the scoring sessions. The tables were prepared for each session separately. Besides making the tables for holistic and analytic scores, I formed tables for the categories of analytic scoring.

Statistical Techniques for Data Analysis

In the analysis of data, I referred to The Research Manual by Hatch and Lazaraton (1991) for information about reliability and interrater reliability. I chose a Pearson correlation matrix to find the correlation coefficients between the raters. Pearson correlation searches for the degree of relationship between two variables. The correlation coefficient is symbolized by the letter "r". The value r is always somewhere between -1 and 0 or 0 and +1. The closer the r is to +1, the stronger the relationship between the variables.

Since Pearson R gives us the reliability for half of the test, I used the Spearman-Brown Prophecy formula to find interrater and intrarater reliability. This formula determines the reliability of the full test. These correlation formulas are used with interval data, so they are appropriate for this study.

CHAPTER 4 RESULTS AND DISCUSSION

Introduction

This study has both tested the reliability of the scoring system of writing section used at Çukurova University (ÇU) and compared the reliability of holistic and analytic scoring methods for essays. The former was described with the obtained observational data during the scoring at the university and then the scores were correlated in order to test the reliability of the system. The latter was the correlational study and the scores given for each method were correlated in order to test the interrater and intrarater reliability of each method.

Description of the Writing Test Process at ÇU

I collected observational data to describe the scoring system at ÇU. I observed how the writing test was given to the students, what procedures were followed before rating, during rating, and during spot-checking.

The Test

The test consisted of four sections: listening, grammar, reading, and writing. In the writing section, in which the researcher was interested, the students were asked to write a descriptive essay on the given topic. The topic was to write a character description for someone using the information given. The information consisted of only some adjectives that referred to people characteristics. The students were asked to write the essays in about 100 words. This meant the word limit for the essay was between 90-110 only. They would lose credit if they had more or less words of this limit. The credit for the writing section was 20 points out of 100 and this section was given in the last thirty minutes.

Students were formed in groups of 25 to 30 in classrooms. The undergraduate advanced level students were grouped in six classrooms. I was present to observe only in one of these classrooms. There were 2 teachers and 29 students in this classroom. The teachers informed the students

about the time allotted for writing section before they started writing. The teachers observed the students and answered their questions when a student wanted to learn the meaning of a word.

When the time for writing section ended, four students said they could not finish writing and they needed more time. However, the teachers did not give them extra time but only let them finish the sentence they were handling.

Pre-rating

The administrators formed a group of eleven teachers to rate the writing section. The teachers were selected randomly, so not all the teachers were teachers of writing. Some of them taught other skills in various levels.

The essays were grouped according to the levels. The raters of the writing section decided to rate the beginner and lower-intermediate level students' essays in the morning and upper-intermediate and advanced level students' in the afternoon. There were 150 essays in the advanced level. Since the advanced level students (the participants in this study) sat for the exam in groups of six, the essays were also grouped into six sets each of which included 25-30 essays. Each rater was supposed to rate ten to fifteen essays for this level.

The Rating Session

The rating session took place one day after the exam in one of the classrooms. The raters came together with a table leader who was one of the testing committee members. There were seven testing committee members formed by the teacher responsible for the testing office and the teachers chosen by the administrators. The table leader gave information about the rubric and answered the questions posed by the raters. The table leader used an OHP to explain the rubric which was a kind of analytic scoring guide (see Appendix B). The rubric had four categories: grammar, coher-

ence & organization, vocabulary, and content & style. The table leader explained the items in general, mostly by reading. No practice session for training took place.

The table leader informed the raters about the rating procedure. They rated the essays set by set (six sets) and divided the rating period into three phases. Two sets of essays were rated in each phase. The first two sets were distributed to the raters. Each rater had 4-5 essays.

The table leader put the rubric for the grammar category on OHP and asked the teachers to first rate the essays according to this category only. She did the same for the other categories when the raters finished rating this category.

During the rating session a discussion arose about scoring according to the number of the words of in an essay. The rubric indicated that students would lose points if they used more or less words than they were asked to write. Some of the raters disagreed with the rubric on this point while some agreed. The discussion took about fifteen minutes. Then the table leader went to the testing office to report the discussion. The testing office responded that they should follow the rubric. Thus, the discussion ended and the raters continued rating following the rubric. The raters wrote the scores for each category and the total score for each essay on essay papers. They then put their initials on these papers.

The table leader collected the first two sets and distributed the other two after they finished rating the first two sets. The table leader put the grammar category on OHP and wrote the rubric for the other categories on the board for this second phase and the third phase. The raters said this was more practical. At the end of these three phases, each rater rated about 12-14 essays.

I also observed the time allotted for the rating of essays. The entire session to rate 150 essays took about 1.5 hours, including the explanation

of the rubric and discussions. I observed that rating took about 3.5 to 4 minutes per essay.

Spot-checking

The day after the rating, the testing committee took the six sets of essays to spot-check. They randomly selected about 25% of the essays in each set. The testing committee rerated the essays that they selected. While rerating, they met discrepancies between the scores given the previous day and the scores they gave in three out of the six sets. Without any correction on the scores, these three sets were given back to the same raters who scored beforehand. These raters rerated the essays. The scores given on the previous day were marked on the essay papers. The raters saw the first scores before their second rating. They reported the changes on the scores to the testing committee after they rerated the three sets.

The other three sets in which the testing committee did not meet any discrepancies between the scores (in 25% of the essays) were not rerated.

Reliability of ÇU Scoring System

Analysis of data revealed that there was a very significant correlation ($r = .97$, $p < .01$) between the scores given before and after spot-checking. Since the closer the r is to 1.00, the stronger the relationship between the variables (scores before and after spot-checking), and since this study accepted .80 as a high reliability, this result shows that the relationship between the raters was very high.

However, it might be a mistake to believe this result. Some sets of essays were rated only once because the testing committee did not meet any discrepancies between the scores while they were spot-checking 25% of these sets. They might have met more discrepancies if all the essays were rerated. Even after the essays were selected for rerating, the second rater saw the first score which was written on the essay paper. This could have affected or biased the second rater.

Interrater Reliability

One hundred and fifty students whose essays were scored at ÇU wrote a second essay for this study. Fifty out of these essays were selected randomly. The selected essays were rated holistically and analytically to find first the interrater reliabilities of these scoring methods.

Interrater Reliability of Holistic Scoring Method

Fifty essays selected by random sampling were rated holistically twice within a month. In both ratings, the scores given by the four raters were correlated and the correlation coefficients and interrater reliabilities were calculated.

Analysis of these holistic scores revealed that there was a significant correlation ($p<.01$) between the raters. The correlation coefficients ranged from .42 to .63 for the first holistic scoring and from .56 to .71 for the second (see Table 1).

Table 1

First & Second Holistic Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.42	.53	.53
Rater 2		1.00	.58	.63
Rater 3			1.00	.57
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.59	.66	.71
Rater 2		1.00	.56	.57
Rater 3			1.00	.64
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

The overall interrater reliability of the two ratings was also quite high, r=.85 (the first scoring .83; the second one .86). This shows that the four raters overlap to the extent of .722 (r^2) while rating holistically. This is considered to be a very high correlation.

Interrater Reliability of Analytic Scoring Method

The fifty essays that were rated holistically were also rated analytically by the same raters twice within a month. In both analytic ratings, the scores given by the raters were correlated.

Analysis of these analytic scores revealed that there was a significant correlation ($p<.01$) between most of the raters. Only the correlation between the third and the fourth raters in the second rating was not significant. Table 2 presents these correlations.

Table 2

First & Second Analytic Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.61	.75	.67
Rater 2		1.00	.66	.53
Rater 3			1.00	.70
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.65	.47	.50
Rater 2		1.00	.45	.58
Rater 3			1.00	.25*
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

* statistically not significant

The overall interrater reliability of the two ratings was very high, .84 (first scoring .88; second scoring .79) as it was in the holistic rating. This result indicated that the four raters overlap to the extent of .705 (r^2) which is a very high correlation. However, the first scoring had a very high reliability whereas the second was low. There was inconsistency between the two scorings.

Interpretation of analysis.

Analysis of holistic and analytic scoring data rejected holistic scoring method had significantly greater reliability than analytic scoring method. Both scoring methods were reasonably high (over .80). However, holistic scoring had consistent interrater reliabilities in both the first and the second ratings (.83 and .86) whereas analytic scoring was inconsistent (.88 and .79). Thus, we can conclude that the analytic scoring method, although its reliabilities are high, does not have as consistent reliabilities as holistic scoring does.

Furthermore, the study considered the training and rating sessions of both scoring methods and found that analytic scoring is much more time-consuming than holistic scoring (for rating: holistic, 1.5 minutes per essay; analytic, 3.5 minutes per essay, and for training: holistic, 1.5 hours; analytic 2.5 hours).

Interrater Reliability of the Individual Categories of Analytic Scoring

The analytic scoring rubric contained five different categories: content, organization, vocabulary, language use and mechanics. Each category was analyzed individually to find whether the reliability of the categories was different from the total scores for analytic rating.

Interrater reliability of content category.

The raters scored the content category while rating the essays analytically. Analysis of data showed that the coefficients of this category between the raters are not very high. The coefficients ranged from .37 to

.58 for the first rating session, and from .13 to .53 for the second (see Table 3). In the first rating there were significant correlations between the raters ($p<.01$). However, in the second rating this significance was found only with two correlations.

Table 3

First & Second Content Category Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.56	.51	.57
Rater 2		1.00	.58	.38
Rater 3			1.00	.37
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.49	.13*	.53
Rater 2		1.00	.33*	.36*
Rater 3			1.00	.24*
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

* statistically not significant.

The overall interrater reliability of the two analytically scored content category was .75 (first rating .80; the second rating .68). The four raters in this category overlap to the extent of .562 (r^2) which can be considered a low reliability.

Interrater reliability of the organization category.

The scores that the raters gave for the organization category while analytic scoring sessions were correlated. Analysis of data for the first and the second organization category scores showed that the correlation

coefficients were quite low, ranging from -.03 to .56 for the first rating and from .26 to .73 for the second. Table 4 presents these coefficients. In this category the correlations between the raters were inconsistent and very low. They were lower in the first rating, including negative correlations. Further, in the first rating, significant correlations ($p<.01$) were found only in the three of the coefficients and in the second rating, four of the coefficients were significant.

Table 4

First and Second Organization Category Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.56	-.16*	.50
Rater 2		1.00	.06*	.51
Rater 3			1.00	-.03*
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.65	.26*	.46
Rater 2		1.00	.37	.73
Rater 3			1.00	.33*
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

* statistically not significant.

The overall interrater reliability of the two ratings for the organization category was .69 (first rating .58; the second rating .79). This result shows that the four raters overlap to the extent of .471 (r^2) which is a very low reliability.

Interrater reliability of the vocabulary category.

The four raters' scores for the analytically scored vocabulary category were correlated. Analysis of data revealed that there was a significant correlation ($p<.01$) between the raters for this category. The correlation coefficients were high in the two rating sessions (see Table 5) when compared to the other categories. The coefficients ranged from .36 to .60 for the first rating, and from .44 to .61 for the second rating.

Table 5

First and Second Vocabulary Category Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.51	.60	.47
Rater 2		1.00	.58	.36
Rater 3			1.00	.38
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.61	.47	.57
Rater 2		1.00	.47	.57
Rater 3			1.00	.44
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

The overall interrater reliability of the two ratings for analytically scored vocabulary category was .80 (the first rating .79; the second rating .81). This shows that the four raters overlap to the extent of .640 (r^2) which can be considered a high reliability.

Interrater reliability of language use category.

The four raters' scores for analytically scored language use category were correlated. Analysis of data showed that there was a significant correlation ($p<.01$) between the raters both for the first and the second ratings. The correlation coefficients (ranged from .35 to .64 for the first rating, and from .37 to .67 for the second rating) were not consistent in both ratings. Table 6 presents the correlation coefficients of this category of the two ratings.

Table 6

First & Second Language Use Category Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.62	.64	.47
Rater 2		1.00	.61	.35
Rater 3			1.00	.54
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.64	.67	.37
Rater 2		1.00	.54	.49
Rater 3			1.00	.38
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

The overall interrater reliability for the language use category in the first and the second ratings was .82 (the first rating .83; the second rating .81). This shows that the four raters overlap to the extent of .672 (r^2) which is a high reliability.

Interrater reliability of mechanics category.

The scores for the analytically rated mechanics category were correlated. Analysis of data showed that the correlation coefficients (see Table 7) between the raters were low, ranging from .19 to .52 for the first rating and from .16 to .57 for the second. Only three of the coefficients in each of the two rating sessions were significant ($p<.01$).

Table 7

First & Second Mechanics Category Correlations of the Second Essays

	Rater 1	Rater 2	Rater 3	Rater 4
FIRST RATING				
Rater 1	1.00	.34*	.46	.19*
Rater 2		1.00	.47	.20*
Rater 3			1.00	.52
Rater 4				1.00
SECOND RATING				
Rater 1	1.00	.56	.57	.25*
Rater 2		1.00	.42	.32*
Rater 3			1.00	.16*
Rater 4				1.00

Note. Z-Transformation was used to average the reliability coefficients.

* statistically not significant.

The overall interrater reliability of the first and the second analytically scored mechanics category was .71 (the first rating .70; the second rating .71). This shows that the four raters overlap to the extent of .504 (r^2) which is a very low reliability.

Interpretation of data together with the analytic categories

Analysis of all data obtained from the holistic ratings, the total scores of the analytic ratings, and the scores given in the analytic categories revealed that a holistic scoring method is more reliable than an analytic scoring method. When we compared the holistic scores and the total scores for the analytic method alone, this difference was not realized. Both methods looked as if they had equal reliabilities. However, analysis of analytic categories showed that this was not the case. Table 8 presents all the inconsistencies and the low reliabilities of these categories when compared to the total score of analytic rating.

Table 8

Interrater Reliabilities of Holistic, Analytic, and Analytic Categories

	First Rating		Second Rating		Average of the two	
	r	r^2	r	r^2	r	r^2
HOLISTIC	.83	.68	.86	.73	.85	.72
ANALYTIC	.88	.77	.79	.62	.84	.70
Content	.80	.64	.68	.46	.75	.56
Organization	.58	.33	.79	.62	.69	.47
Vocabulary	.79	.62	.81	.65	.80	.64
Language Use	.83	.68	.81	.65	.82	.67
Mechanics	.70	.49	.71	.50	.71	.50

According to this table both holistic and analytic scoring methods have high interrater correlations. There is an increase in the second holistic rating, but a decrease in the second analytic rating. This shows that the more the raters are trained on a holistic rubric, the more reliable they become since the raters were more familiar with the rubric in the second rating session. However, this is not true for the analytic rubric. Although the raters were trained again in the second rating session, and

although they were more familiar with the analytic rubric in this session, the reliability was lower.

Table 8 shows the inconsistency of the correlations in the categories. Although the first analytic rating reliability was quite high (.88), the categories did not have such high reliabilities. Some of them were very low, such as organization (.58), and mechanics (.70). On the other hand, the second analytic rating reliability was lower than the first one (.79) but some of the categories, such as vocabulary and language use (.81), were higher than the total score reliability.

Since analytic scoring method is mostly preferred in order to tell students which features in writing they need to improve (rather than to give an overall assessment of writing), the results in this study show that this method might not be effective. The essays that were rated were the same, the raters were the same, and the rubric was the same. However, the results were different. Therefore, I believe that analytic scoring does not give as high interrater reliabilities as holistic does.

Intrarater Reliability

In this study, each rater rated the same essays twice using the same method of scoring. Therefore, the correlations between the first and the second rating were also examined to find their intrarater reliability.

Intrarater Reliability of Holistic Scoring

The scores given in the first and in the second holistic rating by each rater were correlated and then adjusted by the Spearman-Brown formula (SB) to find intrarater reliability of holistic scoring. According to this formula intrarater reliability was .85 for the first rater, .70 for the second, .84 for the third, and .79 for the fourth. Table 9 presents the coefficients and the intrarater reliabilities.

Table 9 shows that the intrarater reliabilities for each rater are reasonably high. Only the second rater shows low correlation although it

is statistically significant ($p<.01$).

Table 9

Correlation Coefficients of Holistic Scoring for Intrarater Reliability

	r	r (SB)	p value
Rater 1	.75	.85	p<.001
Rater 2	.54	.70	p<.001
Rater 3	.73	.84	p<.001
Rater 4	.66	.79	p<.001

Intrarater Reliability of Analytic Scoring

The raters rated the essays twice within a one month interval. The scores for each rater in the first and the second ratings were correlated. According to the SB formula the intrarater reliability was .86 for the first rater, .80 for the second rater, .65 for the third rater and .76 for the fourth rater. Table 10 presents the coefficients and intrarater reliabilities.

Table 10

Correlation Coefficients of Analytic Scoring for Intrarater Reliability

	r	r (SB)	p value
Rater 1	.76	.86	p<.001
Rater 2	.67	.80	p<.001
Rater 3	.49	.65	p<.001
Rater 4	.62	.76	p<.001

The intrarater reliability of analytic rating was lower than holistic. Only two of the raters had high reliabilities.

Since analytic scoring contains five categories, the intrarater reliability of these categories was also examined.

Intrarater reliability of content category.

Each rater's first and second scores for content category of analytic scoring were correlated and then adjusted to SB to find intrarater reliabilities of this category. Intrarater reliability was .83 for the first rater, .73 for the second, .34 for the third and .63 for the fourth. The correlations and intrarater reliabilities are presented in Table 11.

Table 11

Content Category Coefficients for Intrarater Reliability

	r	r (SB)	p value
Rater 1	.72	.83	p<.001
Rater 2	.58	.73	p<.001
Rater 3	.21	.34	*
Rater 4	.47	.63	p<.001

Note. * Statistically not significant

This table shows that there is only a high intrarater reliability for the first rater. The other three have low reliabilities. The correlation between the scores of the third rater is not significant.

Intrarater reliability of organization category.

Each rater's first and second scores for the organization category in analytic scoring were correlated and then adjusted to SB. According to this, intrarater reliability was .81 for the first rater, .75 for the second, .23 for the third and .44 for the fourth (see Table 12).

This table shows how low the intrarater reliabilities is in the three raters. It is high only for the first rater. The third rater's reliability is not significant.

Table 12

Organization Category Coefficients for Intrarater Reliability

	r	r (SB)	p value
Rater 1	.69	.81	p<.001
Rater 2	.60	.75	p<.001
Rater 3	.13	.23	*
Rater 4	.29	.44	p<.05

Intrarater reliability of vocabulary category

The first and the second scores of each rater for the vocabulary category were correlated. According to SB, intrarater reliability was .80 for the first rater, .59 for the second, .46 for the third, and .71 for the fourth (see Table 13)

Table 13 shows that, like the organization category, the intrarater reliability was high only for the first rater on vocabulary.

Table 13

Vocabulary Category Coefficients for Intrarater Reliability

	r	r (SB)	p value
Rater 1	.67	.80	p<.001
Rater 2	.42	.59	p<.01
Rater 3	.30	.46	p<.05
Rater 4	.56	.71	p<.001

Intrarater reliability of language use category.

The first and the second scores of each rater for language use category were correlated and then adjusted to SB. Intrarater reliability of this category was .77 for the first rater, .70 for the second, .63 for the third, and .70 for the fourth. Table 14 presents the correlations and intrarater reliabilities for this category.

Table 14

Language Use Category Coefficients for Intrarater Reliability

	r	r (SB)	p value
Rater 1	.63	.77	p<.001
Rater 2	.54	.70	p<.001
Rater 3	.46	.63	p<.001
Rater 4	.54	.70	p<.001

In this language use category none of the intrarater reliabilities are high although there was consistency among the raters.

Intrarater reliability of mechanics category.

The first and the second scores of each rater for mechanics category were correlated. According to the SB the intrarater reliability of mechanics category was .80 for the first rater, .61 for the second, .61 for the third, and .55 for the fourth. Table 15 presents the correlations and the coefficients of this category.

Table 15

Intrarater Reliability of Mechanics Category

	r	r (SB)	p value
Rater 1	.68	.80	p<.001
Rater 2	.44	.61	p<.01
Rater 3	.44	.61	p<.01
Rater 4	.38	.55	p<.01

In this category again, only the first rater had a high intrarater reliability.

Interpretation of analysis for intrarater reliability

Analysis of data revealed that the intrarater reliabilities were similar in both holistic and analytic scoring methods. In both of them two of the raters had high reliability (over .80). However, the categories of analytic scoring have low intrarater reliabilities. Table 16 presents all the reliabilities of holistic, analytic, and analytic categories.

According to this table the intrarater reliability for both scoring

methods is not consistent. However, this inconsistency is wider in analytic scoring (ranged from .70 to .85 in holistic; from .65 to .86 in analytic). On the other hand, the intrarater reliabilities in analytically rated categories are very inconsistent. There are some very low reliabilities (.23) whereas some are quite high (.83).

Table 16

Intrarater Reliabilities of Holistic, Analytic, and Analytic Categories

	Rater 1		Rater 2		Rater 3		Rater 4	
	r	r^2	r	r^2	r	r^2	r	r^2
HOLISTIC	.85	.72	.70	.49	.84	.70	.79	.62
ANALYTIC	.86	.73	.80	.64	.65	.42	.76	.57
Content	.83	.68	.73	.53	.34	.11	.63	.39
Organization	.81	.66	.75	.56	.23	.05	.44	.19
Vocabulary	.80	.64	.59	.34	.46	.21	.71	.50
Language Use	.77	.59	.70	.49	.63	.39	.70	.49
Mechanics	.80	.64	.71	.50	.61	.37	.55	.30

Table 16 shows that only the first rater was consistent within his/her reliabilities. The other three all have inconsistent reliabilities. Furthermore, the second, third, and fourth raters have very low reliabilities in the categories. As a result, it is reasonable to comment that a holistic scoring method has higher interrater and intrarater reliability than an analytic scoring method.

CHAPTER 5 CONCLUSION

Summary

This study tested two hypotheses. The results of the first hypothesis tested the reliability of the method for scoring the writing test used at Çukurova University Preparatory School. A focused descriptive study combined with correlational design was used to test this hypothesis. A total of 150 EFL preparatory students participated in this study. These students wrote two essays. The first one was written in an exam at ÇU and was rated by the teachers of that university on a kind of analytic scoring method. The scores given for these essays were correlated to test the first hypothesis. The researcher also observed the scoring system at ÇU in order to describe the scoring procedure. The observation procedure was divided into four phases: testing, pre-rating, rating, and spot-checking.

Analysis of correlational data rejected the first hypothesis. The reliability of this scoring method was .97 which is considered a very high reliability. However, the observational data and descriptive analysis showed that the correlation of the scores alone would not be sufficient to claim that this system was reliable. The raters who scored the essays for the second time saw the first scores which could have created bias.

The results of the second hypothesis tested the reliability of holistically and analytically scored essays in this educational context. A correlational design was used to test this hypothesis. The same 150 students wrote a second essay one day after they took the exam. Fifty students out of 150 were selected by random sampling as participants for this hypothesis. The essays of these 50 students were rated twice holistically and twice analytically by four raters in order to test interrater and intrarater reliabilities of each scoring method.

Analysis of data revealed that both holistic and analytic scoring methods had high interrater reliabilities (holistic .85, analytic .84).

However, there was a wider range between the first and the second analytic scoring (.88 and .79) than the first and the second holistic scoring (.83 and .86).

Since the analytic scoring method has five categories (content, organization, vocabulary, language use, and mechanics) each of these categories were correlated individually as well as the total. Analysis of analytically scored categories revealed that the interrater reliabilities of the categories are not as high as the total score reliability of analytic scoring. Furthermore, there was a wide range of interrater reliabilities among the categories (ranging form .69 to .82).

The same fifty essays were also used to investigate intrarater reliability of holistic and analytic scoring. Analysis of data showed that holistic scoring (ranging form .70 to .85) had higher intrarater reliability than analytic scoring (ranging form .65 to .86) and that the raters were more consistent in holistic scoring.

Analysis of analytically scored categories also showed that there was a wide range within and between the categories (ranging form .34 to .83, from .23 to .81, from .46 to .80, from .63 to .77, and from .55 to .80 respectively). Intrarater reliability of the categories was much lower than the total of analytic scoring.

Therefore, it is possible to conclude that holistically scored essays are more reliable and more consistent.

Pedagogical Implications

Those of us who teach EFL in Turkish universities and test our students' writing quality need to be informed about the reliability of scoring methods. We might suppose that an analytic rubric lets us observe the development of students' writing charateristics. This conclusion, according to the current research, is unwarranted because analytic scoring does not have high reliabilities in those characteristics. I believe that

we can observe the development of those characteristics of writing in the classroom itself, when the students write only to learn rather than to be tested. We can help them improve their writing by conferencing with students individually, by letting them revise their writing using multiple draft approach until the teacher and the student feel the writing is improved. For evaluation, we can use a holistic rubric to test if the student reflects what s/he has learned in writing since a holistic scoring method evaluates writing quality as a whole unit, combining syntactic and rhetorical dimensions.

This study, especially my observations at Çukurova University, taught me another thing. Some of the teachers were not very content with the rubric they were using. I think, the teachers and administrators who are responsible with testing can come together before rating and discuss about the rubric until they reach a consensus. Otherwise, the raters who do not like some descriptors might be affected while rating.

Another thing I observed was the way the essays were marked. I think the first score which was written on the essay paper causes bias on the second rater. In order to avoid this bias, we can follow another procedure. Neither the first rater nor the second one marks the essay paper, but writes the scores on different sheets. The two raters, then, come together to compare the scores. If they meet a discrepancy between the two scores, these essays are rated again by a third rater, or the average of the two scores is accepted as the score of the essay.

Suggestions for Further Research

My observation of the rating procedures made me think that we need think-aloud protocol studies during rating. The average length of time was 2-4 minutes for each essay in analytic scoring session. However, one of the raters was rating the essays in 4-6 minutes. This rater was reading the rubric again and again and when I asked if there was something he could

not understand, he said he was only trying to be sure. I believe that a think-aloud study would explain this time difference in rating and the thoughts of this rater. It might be possible to learn why he was always late in analytic scoring.

Additionally, a think-aloud study would also reveal the reactions of the raters to the rubrics used for rating. We would discover if the rubric covered all needed information for rating and if the raters had difficulty in rating some of the essays on the view of the rubric.

BIBLIOGRAPHY

Boyd, B. L. (1990). TOEFL Test of Written English (TWE) scoring guide. TESOL Quarterly, 24, 159-163.

Canale, M., Frenette, N., & Bélanger, M. (1988). Evaluation of minority student writing in first and second languages. In J. Fine (Ed.), Second language discourse: A textbook of current research (pp. 147-166). Norwood, NJ: Ablex.

Carey, L. M. (1988). Measuring and evaluating school learning. Boston, MA: Allyn and Bacon.

Carlson, S. B., & Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), Writing assessment: Issues and strategies (pp. 126-152). NY: Longman.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English (TOEFL Research Report 19). Princeton, NJ: Educational Testing Service.

Connor, U. (1990). Linguistic/rhetorical measure for evaluating ESL writing. In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 215-225). Norwood, NJ: Ablex.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper, & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging (pp. 3-31). Urbana, IL: National Council of Teachers of English.

Diederich, P. B. (1974). Measuring growth in English. Urbana, IL: National Council of Teachers of English.

Greenberg, K. (1986). The development and validation of the TOEFL writing test: A discussion of TOEFL research reports 15 and 19. TESOL Quarterly, 20(3), 531 544.

Gregory, K. (1991). More than a decade's highlight? The holistic scoring consensus and the need for change. (ERIC Document Reproduction Service No. ED 328 594).

Hamp-Lyons, L. (1990). Basic concepts. In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 5-15). Norwood, NJ: Ablex.

Hatch, E. & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. NY: Newbury House.

Henning, G. (1990). Issues in evaluating and maintaining an ESL writing assessment program. In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 279-291). Norwood, NJ: Ablex.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? TESOL Quarterly, 18(1), 87-107.

Hughes, A. (1989). Testing for language teachers. NY: Cambridge University Press.

Hughey, J. B., Wormuth, D. R., Hartfiel, V. F., & Jacobs, H. L. (1983). Teaching ESL composition: Principles and techniques. Rowley MA: Newbury House.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. College Composition and Communication, 41(2), 201-213.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Katz, A. (1988). Second language proficiency: Current issues. Washington DC: U.S. Department of Education Office of Educational Research and Improvement. (ERIC Document reproduction Service No. ED 017 539).

Larsen-Freeman, D., & Long, M. H. (1991). An introduction to second language acquisition research. NY: Longman.

Lauer, J. M., & Asher, J. W. (1988). Composition research: Empirical issues. NY: Oxford University Press.

Lloyd-Jones, R. (1987). Test of writing ability. In G. Tate (Ed.), Teaching composition (pp. 155-176). Texas Christian University Press.

Mann, R. (1988). Measuring writing competency. (ERIC Document Reproduction Service No. ED 295 695).

Mitchell, K. & Anderson, J. (1986). Reliability of holistic scoring for the MCAT essay. Educational and Psychological Measurement, 46, 771-775.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. TESOL Quarterly, 17(4), 651-671.

Stansfield, C. W., & Ross, J. (1988). A longterm agenda for the Test of Written English (Research Rep.). Princeton, NJ: Educational Testing service.

Vaughan, C. (1990). Holistic assessment: What goes in the raters' minds? In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 111-125). Norwood, NJ: Ablex.

APPENDIX A

CONSENT FORM (for teachers)

We are asking you to participate in a study to explore how to evaluate essays better. With your permission, your rating for the essays will be used in the research.

Your identity will not be disclosed and there will be no risk in your participation in this study.

Bilkent MA TEFL Student
Şehnaz Şahinkarakaş

MA TEFL Director
Advisor
Dr. Dan J. Tannacito

****

I have read the information on the form and I consent that my rating will be used in the study of writing assessment. I understand that my participation is completely confidential and that I take no risk involved in my participation.

Name (Print) :_____

Signature :_____

Date :_____

CONSENT FORM (for students)

We are asking you to participate in a study to explore how to evaluate essays better. With your permission, the essay you will write will be used in the research.

Your participation in this study is VOLUNTARY. You should not sign this form if you do not wish to participate. All information will be held in strict confidence. No one will know your identity and there will be no risk in your participation in this study. Your scores will not affect your course evaluation.

Bilkent MA TEFL Student
Şehnaz Şahinkarakaş

Ma TEFL Director
Advisor
Dr. Dan J. Tannacito

****

I have read the information of the form and I consent to be a participant in the study of writing assessment. I know that the essay I will write will be used in the study. I understand that my participation is completely confidential and that there is no risk involved in my participation.

Name (Print) :_____

Signature :_____

Date :_____

APPENDIX B

Rubric Used at ÇU for Writing Section

GRAMMAR

5 : No tense errors and parallelism in tense use.
4-3 : Errors in tense but often do not make intelligibility difficult.
2-1 : Even basic structures - tense used with gross inaccuracies; errors make intelligibility difficult.
0 : Unintelligible

COHERENCE & ORGANIZATION

5 : Flows smoothly from one clearly stated idea to another; every fact or detail relates to the topic; is interesting and satisfying.
4-3 : Ideas clear though not well organized; ideas sometimes repeated.
2-1 : Disorganized and illogical (Ideas/Items are not connected).
0 : Shows no ability whatsoever to link ideas/items.

VOCABULARY

5 : Wide range of vocabulary appropriate to topic; does not repeat same words; very minor spelling errors.
4-3 : Vocabulary appropriate but some repetition occur; some inappropriate words that do not affect intelligibility; a few serious spelling errors.
2-1 : Very limited range of vocabulary; too much repetition; very often inappropriate words.
0 : Shows inability

CONTENT & STYLE

5 : Interesting and appropriate response to the topic; covers all the information given on the question paper.
4-3 : Response to the topic adequate; some information given on the question paper left out; limited ability to match style with content.
2-1 : Inadequate response; very little given information used; no evidence of appropriate style.
0 : No ideas related to the topic expressed.

Note: Word limit for each essay is between 90-110. For every 10 missing or extra words, student will lose 1 point.

APPENDIX C

TWE-Test of Written English

6 **Demonstrates clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.**

A paper in this category
- -effectively addresses the writing task,
- -is well organized and well developed,
- -uses clearly appropriate details to support a thesis or illustrate ideas,
- -displays consistent facility in the use of the language,
- -demonstrates syntactic variety and appropriate word choice.

5 **Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors.**

A paper in this category
- -may address some parts of the task more effectively than others,
- -is generally well organized and developed
- -uses details to support a thesis or illustrate an idea
- -displays facility in the use of language
- -demonstrates some syntactic variety and range of vocabulary.

4 **Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.**

A paper in this category
- -addresses the writing topic adequately but may slight parts of the task
- -is adequately organized and developed
- -uses some details to support a thesis or illustrate an idea
- -demonstrated adequate but possibly inconsistent facility with syntax and usage
- -may contain some errors that occasionally obscure meaning

3 **Demonstrates some developing competence in writing but it remains flawed on either the rhetorical or syntactic level, or both.**

A paper in this category may reveal one or more of the following weaknesses:
- -inadequate organization or development
- -inappropriate or insufficient details to support or illustrate generalization
- -a noticeably inappropriate choice of words or word forms
- -an accumulation of errors in sentence structure and/or usage

2 **Suggests incompetence in writing.**

A paper in this category is seriously flawed by one or more of the following weaknesses:
- -serious disorganization or underdevelopment
- -little or no detail, or irrelevant specifics
- -serious and frequent errors in sentence or usage
- -serious problems with focus

1 **Demonstrates incompetence in writing.**

A paper in this category
- -may be incoherent
- -may be undeveloped
- -may contain severe and persistent writing errors

APPENDIX D

ESL COMPOSITION PROFILE

CONTENT

15-14 : **EXCELLENT TO VERY GOOD:** knowledgeable; substantive; thorough development of thesis; relevant to assigned topic.
13-11 : **GOOD TO AVERAGE:** some knowledge of subject; adequate range; limited development of thesis; mostly relevant to topic, but lacks detail.
10-9 : **FAIR TO POOR:** limited knowledge of subject; little substance; inadequate development of topic.
8-7 : **VERY POOR:** does not show knowledge of subject; non-substantive; not pertinent; OR not enough to evaluate.

ORGANIZATION

10-9 : **EXCELLENT TO VERY GOOD:** fluent expression; ideas clearly stated/ supported; succinct; well-organized; logical sequencing; cohesive.
8-7 : **GOOD TO AVERAGE:** somewhat choppy; loosely organized but main ideas stand out; limited support; logical but incomplete sequencing.
6-5 : **FAIR TO POOR:** non-fluent; ideas confused or disconnected; lacks logical sequencing and development.
4-3 : **VERY POOR:** does not communicate; no organization; OR not enough to evaluate.

VOCABULARY

10-9 : **EXCELLENT TO VERY GOOD:** sophisticated range; effective word/idiom choice and usage; word form mastery; appropriate register.
8-7 : **GOOD TO AVERAGE:** adequate range; occasional errors of word/idiom form, choice, usage but meaning not obscured.
6-5 : **FAIR TO POOR:** limited range; frequent errors of word/idiom form, choice, usage; meaning confused or obscured.
4-3 : **VERY POOR:** essentially translation; little knowledge of English vocabulary, idioms, word form; OR not enough to evaluate.

LANGUAGE USE

12-11 : **EXCELLENT TO VERY GOOD:** effective complex constructions; few errors of agreement, tense, number, word order/function, articles, pronouns prepositions.
10-9 : **GOOD TO AVERAGE:** effective but simple constructions; minor problems in complex constructions; several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured.
8-6 : **FAIR TO POOR:** major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions; meaning confused or obscured.
5-3 : **VERY POOR:** virtually no mastery of sentence construction rules; dominated by errors; does not communicate; OR not enough to evaluate.

MECHANICS

3 : **EXCELLENT TO VERY GOOD:** demonstrates mastery of conventions; few errors of spelling, punctuation, capitalization, paragraphing.

2 : **GOOD TO AVERAGE:** occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured.

1 : **FAIR TO POOR:** frequent errors of spelling, punctuation, capitalization, paragraphing; poor handwriting; meaning confused or obscured.

0 : **VERY POOR:** no mastery of conventions; dominated by errors of spelling, punctuation, capitalization, paragraphing; handwriting illegible; OR not enough to evaluate.